SPSS Instructions Booklet 1

For use in Stat1013 and Stat2001 updated March 2021

Introduction:

The booklet you are holding contains step by step instructions for conducting the statistical analysis needed for introductory statistics courses using SPSS.

How do you know which tool to use? Many students of statistics have trouble answering that question even if using the tool itself is straightforward for them. Before choosing an analysis tool ask yourself the following:

What is the Research Question that I am supposed to answer?

How many variables are involved? For each variable, what is the data type?

Once these questions are answered go to the title page of the booklet and look for the right scenario type and proceed to instructions on exploring the data visually and numerically then (if needed) conduct inferential statistics too.

Number of Variables:

For the purposes of our courses all research scenarios presented will be univariate (one variable) or bivariate (two variables). We will practice in class and you will be able to practice using the web-site www.statcat.ca. Be careful not to confuse categories with variables - slow thinking is crucial for success. If I am measuring the number of males and females in a population I am only dealing with one variable – gender – not two variables male/female. Each variable must be distinct, and one of the types laid out below.

Types of Data: There are more than one way of organizing data types. We will use Categorical (sometimes called qualitative) vs. Measurement (sometimes called quantitative)

Categorical data: comes from a variable that is based on classifying the individual member of a population as belonging to one of 2 or more categories; is further broken down as Nominal and Ordinal – these two are used pretty well universally.

Nominal: data that is broken into categories that are named – and where order is irrelevant.
examples: gender, colour of eyes, favourite musical genre, country of birth.

Ordinal: categories form a natural order, but no real sense of the difference between the levels.
examples: first, second, third in a race, difference between first place and second place in a race may be 0.1 seconds or 5 seconds; letter grades, birth order, likert scale on a survey (i.e. strongly disagree, disagree, neither agree nor disagree etc.)

note1: in some disciplines – notably psychology - ordinal data is treated as measurement data

note2: SPSS does not have a designation called categorical data. It simply allows you to label data as nominal, or ordinal.

Measurement data: comes from a variable that measures a characteristic of an individual member of a population; is often broken up as Interval/Ratio in introductory statistics textbooks. We will not be breaking it up at all.
examples: earnings per hour, shoe size, height of one month old tomato seedlings

note1: SPSS calls measurement data Scale.

In mathematical statistics all data is characterized as either **discrete** or **continuous**. All nominal and ordinal data, and some measurement data is discrete. Some measurement data is continuous. Continuous data is that which can take on all possible intervals between two other numbers, but that is not how we live our lives (eg. we measure the length of a room to the nearest centimeter, not to the nearest 0.0000000000000001 cm). This characterization of data is important for mathematicians, but not as important for introductory statistics in the Health Sciences nor for the casual user of SPSS.

Data analysis (elements of): Data analysis comes after data collection and involves 5 distinct tasks that often get mixed up because of the use of technology.

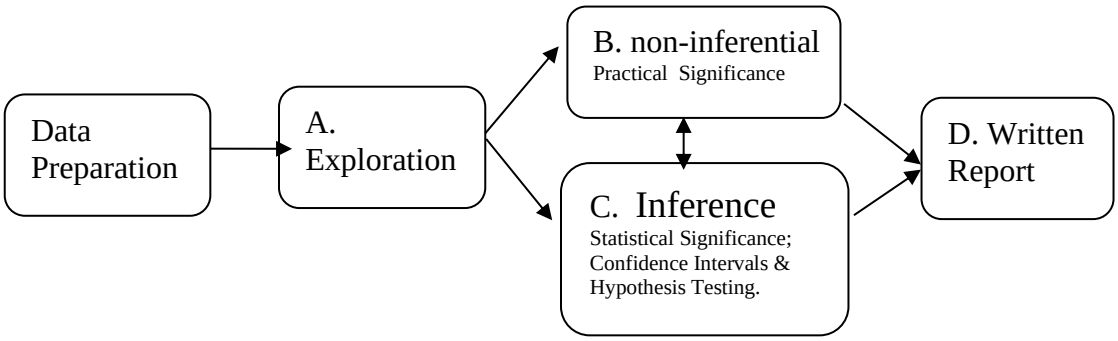


Figure 1: Elements of Data Analysis

Preparation of data: Before analyzing data a statistician needs to make sure that the data is complete, properly formatted and valid, and that he/she understands the context within which the data was collected. This involves communication with the team that collected the data, and a clear understanding of the research question/hypothesis. Activities may include: entering data into electronic format (in our case SPSS), cleaning up data entry errors, and establishing what missing data values may mean. If the data collection and planning aren't done carefully, or if the data is 'messy' then preparation of data can take up a significant amount of time.

Exploration involves a preliminary look (numerical and visual) at the data – simply description is all that is needed here. Taking a look at the distributions of each variable individually (e.g. establishing whether measurement variables are normally distributed or otherwise) precedes looking for relations/associations with two or more variables (i.e. comparing means/proportions or correlation analysis). Exploration may yield a preliminary answer to the research question(s)

Non-inferential (practical significance) involves going one step beyond exploration in order to establish whether there is any relation/association between variables. A descriptive statement like: "The difference between mean diameters of the two types of carrots is 0.5cm." would be the beginning of a search for the strength of relationship. A 0.5cm difference between means of the diameters of two groups of carrot may be meaningful (practically significant), but a 0.5cm mean difference in the heights of two groups of 17 year olds would likely not, thus we need to standardize these measures in some way. You will learn these in stat1013. Practical significance helps establish that there is a story to tell – and is called ***clinical significance*** in medical research. Statisticians often call these and other such statistics measures of effect. Context is key to helping decide whether a finding is 'practically significant', and in many cases the boundary between 'practically significant' and not significant is difficult to pin down.

Inferential – (also called statistical significance) is a process defined by Sally Caldwell as follows: " Inferential statistics is about using sample **statistics** to make inferences (generalizations) about population **parameters**." (e.g. confidence interval for mean) In some research studies 'statistical significance' is incorrectly considered primary to 'practical significance'.

Introduction to the SPSS environment

The two file formats: Raw data and Output

Raw Data file: before and during analysis (file extension .sav) At the bottom of the page you will see two tabs: **Data view**, and **Variable view**.

Data View tab: displays actual data as a spread sheet, columns are variables; rows are separate cases (also known as respondents).

Variable View tab: displays information about each of the variables. Each column is a different aspect of the variable that you can modify.

Output file output of numerical and visual analysis (file extension .spo) can only be opened in SPSS.

Output remains in the output window until you delete it i.e. new output will be added on to previously created output. Be careful to keep track of what output you are working with or delete your output when you are done.

Output can be copied from SPSS and pasted into a Word document for future use. Simply choose the copy command (or cntrl c), you may loose some of the original formatting but you will be able to modify the text in your Word document.

Tip for working with SPSS:

Finding Functions: SPSS is a powerful software package that can do a lot more than most anyone needs. It is made up of many functions/subroutines that have been patched together to make a whole. Tools that seem like they should belong together sometimes do not (eg. Chi-square tests). Also, you can often find more than one function that will do a required task, especially with the exploratory statistics.

Where is the Data? Entering and importing data.

You will have one of four options:

1. Data entry

1. Click **File → New → Data**
2. Go into **Variable View**
3. Write in your 'nameofvariable' under **Name** and SPSS will automatically designate the rest of the fields (columns). (no spaces/periods etc. allowed in the names)
4. For some data you will need to change the **Measure** to the appropriate setting. (remember scale is equivalent to ratio/interval)
5. Before entering nominal/categorical data you will need to code it: For example male/female can be coded as 1/0, urban/suburban/rural can be coded as 1/2/3. To prepare SPSS for this you need to go to the **Values** column and click in the appropriate row. A box will open up, then for each code separately you need to enter the **Value**, and the **Value label**. eg. value = 1, value label = male; value = 2, value label = female
6. Go into **Data View**
7. Enter the data in the appropriate column.
   a) to enter measurement data just enter the measurements in the appropriate column
   b) to enter nominal/ordinal data enter the appropriate value the correct number of times
   eg. 55 males (value male = 1) means you enter a '1' 55 times.
8. Save the data set.
9. You are now ready to do the work on the new data set.

2. Importing data from Excel

- every column in Excel must represent a separate variable
- it is best if you label each column with a variable name in SPSS format (i.e. no blanks etc.)
- import the file by clicking on **File → Open → Data** then selecting excel in the 'file of type' box
- follow the instructions and check your SPSS data file carefully to make sure the import worked perfectly before using it.

Exploration: one measurement variable

Numerical: Calculating the mean, median, and mode, standard deviation/variance, range, min/max values, and quartiles/percentiles:

1. In the Data Editor/Data View window choose **Analyze/Descriptive Statistics/Frequencies**.
2. The **Frequencies** dialogue box opens. Click on the desired variable to highlight it, then click on the right arrow to move it to the variable(s) box. (or you can just double-click on the variable name and it will move to the variable(s) box).
3. Make sure that the *Display frequency tables box* is **NOT** √ as this produces a frequency table, which is not useful here.
4. In the **Frequencies** dialogue box, click on *Statistics…* and the **Frequencies: Statistics** dialogue box will appear. Select the desired statistics. Click continue. Click OK – check output.

Note: you can also get at most of these descriptives using **Analyze/Descriptive Statistics/Descriptives (or Explore)**, but you can only get mode and percentiles through **Frequencies**.

Visual:

The Histogram: There are two ways to get a histogram (neither of which allow you to control the width of the class.

1. In the Data Editor/Data View window, select **Graphs→ Legacy Dialogues →Histogram**. The Histogram dialogue box will appear.
2. In the **Histogram** dialogue box move the desired variable to the Variable area. You will need to double-click or highlight the desired variable and use the appropriate arrow to move it. To display a Normal Curve over top of the Histogram, √ this box. Click OK – check output

The Boxplot: box plot displaying one measurement variable.

1. Go to **Graphs→Legacy Dialogues →Boxplot.** That will generate the ***Boxplot*** dialogue box.
2. Select ***Simple***, ***Summaries for separate variables*** and then click on **Define**.
3. In the ***Define Simple Boxplot, Summaries for separate variables*** dialogue box move the measurement variable to the ***Boxes represent*** slot and leave the rest blank; click ok check output

Note: you can also get a boxplot through the **Analyse/Descriptive Statistics/Explore** function

The Dot Plot is available through **Graphs→**Legacy Dialogues

1. Go to **Graphs→Legacy Dialogues →Scatter/DotPlot.**
2. In the ***Scatter/Dot*** dialogue box select 'simple dot' and then click **Define**.
3. move the desired variable to the x-axis slot click ok and check output

Exploration plus: one measurement variable
Percentiles for individual raw scores
calculating z-scores
& Checking for Normality

<u>Percentiles for individual raw scores:</u> These need to be read from the cumulative frequency table produced as follows:

1. In the Data Editor/Data View window choose **Analyze/Descriptive Statistics/Frequencies**.
2. Make sure that the *Display frequency tables box* is **<u>clicked</u>** √ - this produces a frequency table, which includes a 'cumulative frequency column.
3. Look up the desired raw score and the corresponding cumulative frequency will be the percentile

<u>Calculating Z-scores for every case</u>: These will be added to your data set, instead of into an output file.

1. Select **Analyze/Descriptive Statistics/Descriptives**
2. In the main **Descriptives** dialogue box, select the variable(s) you would like to have z-scores for. You cannot calculate a z-score for one case, just for every case in a variable.
3. Check the box that says **Save standardized values as variables**. SPSS will calculate z scores for each of the variables using the formula you learned about and append them to the end of your data file - click **ok**
4. To view the z scores, open up your data set and click on **Data View** (at the bottom left of the screen). You can see that SPSS named each selected variable with a z and then truncated the original variable name. These names also appear in **Variable View**. You may save this data set with another name, if you choose, or delete the variables one by one.

<u>Checking For Normality</u> – Skewness, Kurtosis and visual Normal Curve

1. Check if Mean = Median = Mode through the **Analyze/Descriptive Statistics/ Frequencies** function
2. Skewness and Kurtosis are available through **Analyze → Descriptive Statistics → Descriptives/Frequencies (or Explore)**
3. Check visually for a bell curve shape by plotting a histogram and superimposing a 'perfect' normal curve on top as follows: **Graphs → Legacy dialogues →Histogram** then select **with Normal Curve** after selecting the Histogram.
4. Click **O.K.** check your output.

Exploration: one categorical variable

Numerical: Creating a frequency table without a graph:

1. Make sure all variables are set to display labels rather than the numerical values to make the graphs/ tables more meaningful (go into **View** and selecting **value labels.)**
2. In the Data Editor/Data View window choose **Analyze/Descriptive Statistics/Frequencies**.
3. The **Frequencies** dialogue box opens. Click on the desired variable to highlight it, then click on the right arrow to move it to the variable(s) box. (or you can just double-click on the variable name and it will move to the variable(s) box). Make sure that the *Display frequency tables box* is √ , this will produce a frequency table for the desired variable. Click ok and check your Output.

Visual: Creating a simple pie chart through Graphs function:

1. In the Data Editor/Data View window, select **Graphs → Legacy dialogues →Pie** . The Pie Charts dialogue box will appear.
2. In the **Pie Charts** dialogue box select '*Summaries for groups of cases*' then click on *define*.
3. In the **Define Pie: Summaries of Groups of Cases** dialogue box move the desired variable to the **define slices by** slot. You will also need to select the appropriate representations for the slices (in most cases **Slices Represent no. of cases** will be appropriate) click O.K.
3. Check your Output.

Visual: Creating a frequency bar graph while creating a table:

1. go through steps 1 to 3 above
2. In the **Frequencies** dialogue box, click on *Charts…* and the **Frequencies: Charts** dialogue box will appear. Select the desired chart (bar or pie). Click continue. Click OK.
3. Check your Output.

Visual: Creating a simple frequency bar graph through Graphs function:

1. In the Data Editor/Data View window, select **Graphs → Legacy dialogues →Bar** . The Bar Charts dialogue box will appear.
2. In the **Bar Charts** dialogue box select '*Simple*' and '*Summaries for groups of cases*' then click on *define*.
3. In the **Define Simple Bar:Summaries of Groups of Cases** dialogue box move the desired variable to the **category axis** slot. You will also need to select the appropriate representations for the bars (in most cases **Bars Represent no. of cases** will be appropriate) click O.K. and check your Output.

Note: you can choose frequencies or percentages when creating bar/pie charts. Both will look the same, but the vertical axis will have different values.

Exploration: two variables
Both measurement (Correlation & Regression)

Numerical: Pearson's r - The correlation coefficient,

1. Click **Analyze ⇨Correlate ⇨Bivariate**. The **Bivariate Correlations** dialog box will appear
2. Double-click on the desired variables to move them to the **Variable(s) box**
   The order in which you select the variables influences only the order in which the variables are listed in the output, which does not really matter here.
3. Click **OK** (clicking OK to run the analysis without changing anything else produces
   a. Pearson correlation coefficients,
   b. two-tailed significance test (used for inferential statistics - not for numerical exploration)

Visual: The Scatterplot.

1. Click **Graphs ⇨ Legacy dialogues ⇨ Scatter**
2. Click **Simple**, then click **Define**. The **Simple Scatterplot** dialog box will appear.
3. Click "**dependent variable name**", and move the variable label to the Y-axis box.
4. Click "**independent variable name**", and move the variable label to the X-axis box.
5. Click **OK**, and you will see the scatterplot.

Visual: adding a line of best fit to the scatterplot :

1. Double-click on the **scatterplot** image itself, the chart editor window will open.
2. Click on **ONE SINGLE DOT** only until all points turn a colour - yellow or blue - (this sometimes requires several attempts)
3. At the top of the Chart Editor box, click the **Add fit line at total** icon in the menu bar, and then choose a '**linear**' under 'Fit Method' . Click apply, line of best fit will appear.
4. Click outside the Chart Editor and your scatterplot will have the fit line.

Numerical: The equation of the line of best fit (regression line) (for SPSS 22 and earlier only)

1. Click **Analyze ⇨ Regression ⇨ Linear**. The **Linear Regression** dialog box will appear.
2. Click on the variable named "**dependent variable name**", and move it to the **Dependent** variable box. This is the variable being predicted.
3. Click on the variable named "**independent variable name** ", and move it to the **Independent(s)** variable box.
4. Click **OK** and check output. You will generate a few tables of data:
   a. The **model summary table** gives you r^{2-} which is the **coefficient of determination**.
   b. The **ANOVA** table can be ignored.
   c. The **coefficients** table gives you the slope of the regression line (β) - second row under the column labeled B, and the y-intercept (α) – first row under the column labeled B to generate the equation of the regression line: $Y = \alpha + \beta x$.
   In the example below: $Y = 96.707 + 0.792x$

Coefficients(a)

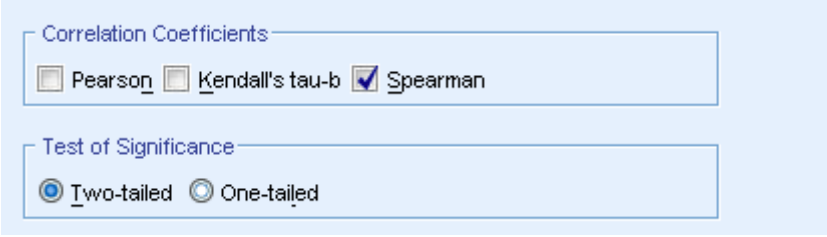
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	96.707	5.389		17.944	.000
	Age in years	.792	.130	.820	6.073	.000

a Dependent Variable: Systolic Blood pressure

Exploration: two variables
Both measurement (special case)

Spearman's rho - The correlation coefficient for two ordinal variables treated as measurement.

1. Click **Analyze ⇨Correlate ⇨Bivariate**. The **Bivariate Correlations** dialog box will appear
2. Double-click on the desired variables to move them to the **Variable(s) box**
   The order in which you select the variables influences only the order in which the variables are listed in the output, which does not really matter here.
3. Under correlation coefficients deselect Pearson and select Spearman's rho



4. Click **OK** and check output which can be analysed in the same way that you would analyse Pearson's r output.

Exploration: two variables
Both categorical (two or more categories)

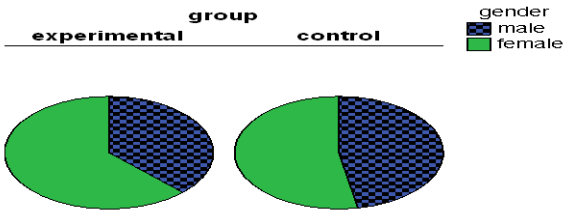
In most circumstances you will first want to explore each variable individually. For example a pie chart, or a frequency table for each variable. You can represent two categorical variables simultaneously through crosstabs (aka contingency table) and visually through a comparison pie chart. Once there are >2 categories it can get very messy. We will focus mostly on scenarios with 2 categories each in both the stat1013 and stat2001 courses.

Numerical: Creating a frequency table with two variables – Crosstabs:

1. In the Data Editor/Data View window, select **Analyse/Descriptives/Crosstabs**. The *Crosstabs* dialogue box will open.
2. In the *Crosstabs* dialogue box move one of the categorical variables (preferably the independent variable) into the *Rows* slot and the other (the dependent variable) into the *Columns* slot. Click: o.k.
3. If you wish to display percentages as well as counts you will have to select ***Cells*** and within the ***Cells dialogue box*** choose ***percents*** by rows or by columns, depending on what you need. Don't select everything or you won't be able to read the chart. (If you have the independent variable in rows and the dependent variable in columns then select rows – this will produce easy comparisons of rates.)

Visual: Paneled pie charts two or more categories:

1. In the Data Editor/Data View window, select **Graphs → Legacy dialogues →Pie** . The Pie Charts dialogue box will appear.
2. In the **Pie Charts** dialogue box select '*Summaries for groups of cases*' then click on *define*.
3. In the **Define Pie: Summaries of Groups of Cases** dialogue box move the desired variable to the **define slices by** slot. You will also need to select the appropriate representations for the slices (in most cases **Slices Represent no. of cases** will be appropriate) click O.K.
3. There is another dialogue box titled ***Panel by:*** This allows you to generate one pie chart for each category in the independent variable. For example: an experiment was conducted (i.e. there was a treatment/experimental group and a control group). You would like to see if the distribution of gender was the same in the control vs. the experimental groups. Thus you are investigating two variables: Gender (male/female) and Group (experimental/control). Plotting pie charts (paneled by column) generates the following graph. You need to be careful which variable to put into the 'paneled by' box. Use trial and error at first – eventually you will discover that the independent variable goes in 'paneled by'.



Exploration: two variables
One Categorical One Measurement.

Numerical: Calculating the mean, median, standard deviation/variance, range, min/max values, for each category of the categorical variable:

1. Analyse/Compare Means/Means then the 'Means' dialogue box will open.
2. Move the categorical variable to the Independent list and measurement variable to the dependent variable list.
3. Click on options and choose the descriptive statistics you want in your table (means, std. dev. Etc.)
4. Click on continue and ok – and check your output.

Visual: Graphs can be very useful in going beyond simply comparing means. With them you will be able to compare the distribution of the measurement variable by each of the categories of the categorical variable and look for outliers.

Comparison Boxplot: box plot displaying one measurement variable.

1. Go to **Graphs→Legacy Dialogues →Boxplot.** That will generate the ***Boxplot*** dialogue box.
2. Select ***Simple***, ***Summaries for groups of cases*** and then click on **Define**.
3. In the ***Define Simple Boxplot, Summaries for groups of cases*** dialogue box move the measurement variable to the ***Boxes represent*** slot and and the categorical variable to the **category axis** slot; click ok check output

Paneled histogram: two or more separate histograms

1. In the Data Editor/Data View window, select **Graphs→ Legacy Dialogues →Histogram**. The Histogram dialogue box will appear.
2. In the **Histogram** dialogue box move the desired variable to the Variable area. You will need to double-click or highlight the desired variable and use the appropriate arrow to move it. To display a Normal Curve over top of the Histogram, √ this box. Click OK – check output
3. In all graphs functions there is a dialogue box titled ***Panel by:*** This allows you to generate one histogram for each category in the second variable and lay them out in one row or one column.

Inferential: one variable
Measurement
Confidence Intervals and hypothesis test of one mean

Generating the <u>Confidence interval</u> for one mean:

.

1. Select **Analyze/Descriptive Statistics/Explore**
2. Move <u>all</u> the desired variables to the **Dependent List** box.
3. Under **Display** select **statistics** only as we don't need plots this time**.** Within Statistics select **Descriptives.** Make sure that the Confidence Interval is set at 95%
4. Click **Continue/O.K.** and check your output.
5. You will get a whole list of Descriptive Statistics, the Confidence Interval is designated between the ***lower bound*** and the ***upper bound***.

Generating the p-value in a <u>hypothesis test for one mean</u>

1. Choose **Analyze/Compare Means/One Sample T Test** tool. This produces the ***One Sample T Test*** dialogue box

2. Move ***the desired variable*** to the test variable box. Lower down you will see a box labeled **Test Value**. Enter the estimated/predicted value of μ from your null hypothesis test (H_0).

3. Under Options choose the confidence level. (usually 95%) Click OK. Check your output.

Inferential: one variable
Categorical: Chi-Square Goodness of Fit Test
(+ Hypothesis test 1 proportion)

χ^2 'goodness of fit test' if **H_0: *all categories equal***

1. Click **Analyze/Nonparametric Tests/Chi-Square**. The **Chi-Square Test** dialog box will appear.
2. Double-click on the variable of interest to move it to the **Test Variable List** box.
3. Under **Expected Values**, choose "All categories equal" (default).
4. Click **OK** and take note of the relevant output (i.e. the calculated χ^2 statistic, and the *p*-value which is called asymp. sig.)

χ^2 'goodness of fit test' if **H_0: *all categories are not equal***

1. Investigate the categories of the variable in question and decide expectations (in terms of percentage) for frequency in each category. (note: the sum for all categories must be 1)
2. Click **Analyze/Nonparametric Tests/Chi-Square**. The **Chi-Square Test** dialog box will appear.
3. Double-click on the variable named "**exercise**" to move it to the **Test Variable List** box.
4. Under **Expected Values**, choose "values", and then list the expected proportions of all categories. If you are not sure of all of the categories, look into the values column for the chosen variable in variable view. Beware of missing values – they come up as 99 in the output.
5. Click **OK** and take note of the relevant output (i.e. the calculated χ^2 statistic, and the *p*-value which is called asymp. sig.)

Hypothesis test 1 proportion: 2 categories only – not necessary as the $\chi 2$ 'goodness of fit test' is sufficient and equivalent

Using SPSS to compute **z-test for single sample proportion test.** Follow the steps below.

1. Open the Data Set: Click **Analyze/Non-Parametric Tests/Binomial**. The **Binomial Test** dialog box will appear.
2. In this dialog box, your list of variables appears in the box to the left. You must choose one or more of the variables and move it/them to the 'Test Variable List'. This is the variable (or variables) for whom you wish to test the proportion of 'successes'.
3. Next, you set the desired 'Test Proportion' to a value between 0.001 and 0.999
4. Click **OK,** Check your output. This should give you a straightforward chart with the significance, and the z_{calc} values

Write out an appropriate final statement.

Confidence interval for proportion must be calculated by hand (or by SPSS script, or by using Excel etc.) as there is no built in function to do this in SPSS.

Inferential: two variables both measurement hypothesis test of Pearson's r

The SPSS function is set appropriately to test whether the value of 'r' is significantly different from zero. i.e. H_0: r = 0.

In order to run this on SPSS you need not do anything as the two tail test is the default that is calculated every time you run a Correlation to get Pearson's r.

Hypothesis test of Pearson's r - The correlation coefficient,

1. Click **Analyze ⇨Correlate ⇨Bivariate**. The **Bivariate Correlations** dialog box will appear
2. Double-click on the desired variables to move them to the **Variable(s) box**
   The order in which you select the variables influences only the order in which the variables are listed in the output, which does not really matter here.
3. Click **OK** (clicking OK to run the analysis without changing anything else produces
   a. Pearson correlation coefficients,
   b. two-tailed significance test (used for inferential statistics - not for numerical exploration)

In the sample output below the Pearson r value is 0.006 and the *p*-value is 0.979.

Correlations

		test1	test2
test1	Pearson Correlation	1	.006
	Sig. (2-tailed)		.979
	N	25	25
test2	Pearson Correlation	.006	1
	Sig. (2-tailed)	.979	
	N	25	25

Inferential: two variables both categorical
χ^2 Test of Independence

To compute a **χ^2 'test of independence'**,

1. Open a Data Set
2. Click **Analyze/Descriptive Statistics/Crosstabs**. The **Crosstabs** dialog box will appear.
3. Move the names of the variables you wish to analyze into the appropriate boxes on the right. The choice of row versus column variables is arbitrary. It doesn't matter which variable is the row variable and which variable is the column variable.
4. **Don't** click OK just yet.
5. Click on the **Cells** button at the bottom of the dialog box. This produces a new dialog box. To select data to be displayed in your contingency table under **Counts**, select **Observed** and **Expected**. Under **Percentages**, select one of **Row**, **Column**, or **Total,** depending on what you need. Click **Continue**.
6. At the bottom of the main **Crosstabs** dialog box, click on **statistics** and select **Chi-square**.
7. Click on **Continue** to return to the main **Crosstabs** dialog box
8. Click on **OK**
9. Write out the relevant output (i.e. the calculated χ^2 statistic, and the *p*-value)

Hypothesis test with 2 proportions: In many introductory statistics textbooks and courses the hypothesis test of 2 proportions is taught. SPSS does not have a function to carry this out, but conducting the **χ^2 Test of Independence** will generate the same results as for the 2 proportions test in most circumstances.

Inferential: two variables
One Measurement and One Categorical (≥2 categories)
ANOVA and Post Hoc

To compute an **ANOVA**, follow the steps below

1. Click **Analyze/Compare Means/One-Way ANOVA**. The **One-Way ANOVA** dialog box will appear.
2. In this dialog box, your list of variables appears in the box in the upper left. Move the measurement variable to the **Dependent List** box. Move categorical variable to the **Factor** box.
3. In the One-Way ANOVA dialogue box, click **Options**
4. In this next dialogue box, click **Descriptive** under **Statistics;**
5. The default under **Missing Values** is **Exclude cases analysis by analysis**. This is fine. click **Continue;** click **OK;**
6. check your output

To compute an POST HOC Test to determine which pairs of means are statistically significantly different (only do this if you have rejected the null hypothesis with the ANOVA test).

1. Click **Analyze / Compare Means/ One-Way ANOVA**. The **One-Way ANOVA** dialog box will appear.
2. In this dialog box, your list of variables appears in the box in the upper left. Move the measurement variable to the **Dependent List** box. Move the categorical variable to the **Factor** box as it is the independent variable. Most often this variable will be labeled "group".
3. In the one way ANOVA box click on the **Post Hoc** button. You will be provided with a list of Post Hoc tests. Each uses a slightly different approach to comparing the differences between means. Under the equal variances assumed heading choose **Tukey** (honestly significant differences).
6. Click **Continue;** click **OK;** check your output.

Inferential: two variables – One Measurement and One Categorical (2 categories) hypothesis test 2 means (independent)

To compute **t-test for independent groups**, follow the steps below.

1. Open the Data Set
2. Look at the numerical codes for the categorical variable (the groups within the variable).
3. In the Data Editor/Data View window choose **Analyze ⇨ Compare Means ⇨ Independent-Samples T Test**. The **Independent-Samples T Test** dialog box will appear.
4. Move one or more ***desired variables*** into the box labeled **Test Variable(s)** to select your dependent variable(s). Then you must move one of your variables into the box labeled **Grouping Variable** to identify the groups to be compared. This will be your independent variable.
5. When you select your independent variable (called **Grouping Variable)**, the button labeled **Define Groups** becomes functional. Click on this button and the **Define Groups** dialog box appears. You must specify the two values of **group** that represent the two groups you wish to compare.
6. Click on **Continue** to return to the **Independent-Samples T Test** dialog box.
7. Click **OK**
8. Check your output

If you need Confidence Intervals for the mean (by category) (+ graphs) follow the steps below:

1. In the **Analyze/Descriptive Statistics/Explore** and the Explore Dialogue box will open.
   Move the measurement variable to the **Dependents list** slot and the categorical variable to the **Factor list** slot.
2. Click **Statistics** and choose the desired statistics, then **continue.**
3. Under **Display** select **both** or only **statistics** if you do not want any plots.
4. Click **Plots** and choose the desired plots, then **continue & O.K.** and examine the output**.**

Random Sampling – a taste of Inferential Statistics:

Setting the random number generator 'seed': In order to set your own random number 'seed' do the following

1. Go to **Transform/Random Number Generators**

2. In the **Random Number Generators** dialogue box click in the box beside **Set Active Generator** select the **Mersenne Twister**, then click on the **'Set Starting Point'** box, and select **'Fixed Value'**

3. In the empty box beside Value: choose a number – 5 or more digits in length - in order to ensure that your number is more likely to be unique. Click OK to enter your value as the 'seed'.

Taking a Random sample (size n) from a data set (size N). You will be creating a new data file with only the sampled data in it.

1. Go to **Data/Select Cases.** That will generate the ***Select Cases*** dialogue box.
2. Under ***Select*** Select **Random Sample of cases** and click on **Sample.** This will generate the ***Select Cases:Random Sample*** dialogue box

3. Select ***Exactly*** 'n' ***cases from the first*** 'N' ***cases***. click on Continue.

4. Under ***Output*** Select **Copy selected cases to a new data set** and enter a variable name for the sample **.**

5.. Click o.k. and check your output.

Transforming variables.

Often the raw data that one has in front of oneself needs to be manipulated in order to be more useful. Here we will look at two such circumstances.

1. Recoding categorical variables which has many categories into 2 categories. We will start with an assumption that there are 5 categories (strongly disagree – 1, disagree – 2, neither disagree nor agree – 3, agree – 4, strongly agree – 5). Assume that the researcher is really only interested in those who either agree/strongly agree (1) vs the rest(0)
   Through Transform/Compute Variable open up the **Recode into Different Variables** dialogue box**.**
   First: Move the variable in question into the middle 'window' and write in a name for the new variable: e.g. satisfaction_recoded
   Second: click on Old and New Values, and in the 'Old and New Values' dialogue box give each 'old' value a 'new' value. There are various ways of doing this, but in the end Old values (1,2,3)= (0) in the new and (4,5) = (1) Click Continue, OK and your new variable is ready to go.

2. Transforming the scores of a variable or set of variables into another variable
   a. E,g, variable 'test_score' comes out of 57, and you need to convert to %.
      Through Transform/Compute Variable open up the **Compute Variable** dialogue box**.**
      First: write in a name for your target variable (In this case test_score_percent makes sense)
      Second: Take the original variable (appearing in a column to the left of the dialogue box) and shift it into the numeric expression box with appropriate arithmetical operators. In this case it could read test_score*100/57.

   b. 5 survey questions (variables) Q1, Q2, Q3, Q4, Q5 each ask about agreement with regards to a classroom experience with scores between 1 and 5 (strongly disagree to strongly agree). You want a total score for the survey (or part of a survey)
      Through Transform/Compute Variable open up the **Compute Variable** dialogue box**.**
      First: write in a name for your target variable (In this case sum_of_scores makes sense)
      Second: create an equation in the numeric expression box with appropriate arithmetical operators and variables. In this case it would read Q1+Q2+Q3+Q4+Q5. Click Continue, OK and your new variable is ready to go.

General tips on editing a graph once it is created:

1. Decide on what changes need to be made (colours etc.)
2. To edit a graph, double-click on the graph in Output View and the SPSS Chart Editor box will open up. There are many options which will not be explained in detail here – you will need to play around with this editor on your own.
3. Make the edits that you see fit. Keep them simple until you understand what you are doing. Click on apply once you are done with each individual change.
4. To return to the original graph, close the Chart Editor by clicking on the icon in the very upper left hand corner of the window.
5. To copy the graph into a word document simply click on it once to select it then copy (ctrl c) and past (ctrl v) into the word (or other) document. You will not be able to change any attributes except size the 'picture' leaves the SPSS platform.

Exploring 2 variables where one is time: graphing timelines

The principle is based on the fact that each case is a separate – evenly spaced - time period or consecutive event (e.g. year, day, month or 1st 2nd 3rd etc.) If so, follow the following steps

1. Click **Graphs/Legacy Dialogues/Line** The **Line Charts** dialogue box opens**,**

2. Select **'Values of individual cases'.**

3. Move the variable you wish to plot into the 'Line Represents' box and click OK.

4. Examine the output, 'Case Number' in fact represents the consecutive time periods/events.

Special case – not used in our courses, just FYI.

There is another option for analyzing 2 measurement variables, but only if the cases are paired as pre/post tests or related persons father/son or siblings it is called the paired means test – and I present it in class as a single means analysis, but it is presented in other textbooks and courses as a paired means test.

To compute inferential **t-test for paired (related) groups**, follow the steps below.

1. Open the Data Set
2. In the Data Editor/Data View window choose **Analyze ⇨ Compare Means ⇨ Paired-Samples T Test**. The **Paired-Samples T Test** dialog box will appear.
3. Move both of the ***desired variables*** into the box labeled **Paired Variable(s)**.
4. Click on **Continue** to return to the **Independent-Samples T Test** dialog box.
5. Click **OK**
6. Check your output – you will see output for both the correlations test from the previous page and the paired t-test as well. If there is no difference between the pairs, you will see that they are highly correlated, which is what you should expect (this may need a dark room moment).